DOCUMENT RESUME

ED 189 165 TM 800 337

AUTHOR Hutten, Leah R.
TITLE Some Empirical Evidence for Latent Trait Model Selection.
PUB DATE Apr 80
NOTE 33p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Factor Structure; *Goodness of Fit; Guessing (Tests); *Latent Trait Theory; Multiple Choice Tests; Scores; Scoring
IDENTIFIERS *Birnbaum 3 Parameter Logistic Model; Distributions (Statistics); *Rasch Model

ABSTRACT

The results of this study suggest that for purposes of estimating ability by latent trait methods, the Rasch model compares favorably with the three-parameter logistic model. Using estimated parameters to make predictions about 25 actual number-correct score distributions with samples of 1,000 cases each, those predicted by the Rasch model fit the observed distribution of number-correct scores as well as the distributions predicted by the three-parameter model. This suggests that the Rasch model is quite robust to violations of its underlying assumptions and can therefore be applied in a wide variety of situations. The single model assumption related to fit was that of unidimensionality. Not only was fit for multidimensional tests quite poor, but it was also difficult to obtain parameter estimates on such tests. Other research indicates that the three-parameter model may be preferable for test development, test equating, and studies of item bias, but for estimating ability, this study shows that the Rasch model produces estimates of ability that rival those produced by the three-parameter model. Further, the study demonstrates that Rasch parameter estimates are obtained at costs significantly reduced rom from those of the three-parameter estimates. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED189165

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SOME EMPIRICAL EVIDENCE FOR LATENT TRAIT MODEL SELECTION[1,2]

Leah R. Hutten[3,4]
Wellesley College

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
L. Hutten

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 800337

---

1 A paper presented at the Annual Meeting of AERA, Boston, April, 1980.

2 This work was originally supported by a grant from the U.S. Air Force Office of Scientific Research.

3 Formerly of the University of Massachusetts, Amherst.

4 Under the guidance of Ronald K. Hambleton.

SOME EMPIRICAL EVIDENCE FOR LATENT TRAIT MODEL SELECTION

Leah R. Hutten
Wellesley College

Latent trait theory has become an accepted alternative to classical test theory for solving measurement problems, but there still remain questions concerning which of this family of models should be selected for various practical applications. Rather than exploring all of the possible latent trait models, this paper focuses on the more general of the latent trait models, the Rasch or one-parameter model, and the three-parameter logistic model. The proponents for each of these two models claim that theirs is the appropriate model for explaining behavior on mental tests. In the case of the three-parameter logistic model, the theory is general enough to encompass most tests provided that they are unidimensional. The Rasch model, on the other hand, lacks some of the general attributes of the three-parameter model. First it assumes that all item discriminations are equal, an assumption that is rather restrictive. Then the model assumes that there is no guessing, an assertion which seems highly inappropriate for frequently used multiple choice tests. But the Rasch model is simpler both in its conceptual basis and utility, so it is tempting to overlook these shortcomings. The Rasch model, as a limited case of the more general three-parameter model, also

assumes that tests are unidimensional.

The purpose of this research has been to compare the "goodness of fit" of the one and three-parameter logistic models with empirical data. The study specifically examined the appropriateness, or fit, of the models to data when the underlying model assumptions were violated. Model fit can be defined both in terms of item fit or person fit. The test developer is interested in the appropriateness of items to a model, whereas the practitioner is primarily concerned with the capability of the model for estimating ability. This research focuses on the question of fit specifically in terms of ability.

Model fit has been an important consideration in latent trait research. Fit statistics are frequently reported in the literature and computer programs devised for latent trait parameter estimation typically include statistics for model fit in their outputs. In examining latent trait research, it seems that both models have "fit" most data to which they've been applied. In most cases the fit of the given model has been sufficient for researchers to assume that the selected model fits empirical data. Unfortunately, one finds few studies which compare fit of latent trait models.

Hambleton (1969) and Hambleton and Traub (1973) compared fit of empirical data to the Rasch and two-parameter models. They found that when item discriminations were

unequal, the two-parameter model displayed improved fit over the Rasch model. Computer programs for obtaining parameter estimates were not available at the time of these studies, so approximations were found. The parameter estimates were then used to predict distributions of number-right scores. A weighted number-right score criterion was used in which item discriminations served as weights (all weights were 1's in the Rasch model case). The research reported herein follows the methodology of Hambleton and Traub, but in the current study unweighted raw scores were used.

Koch and Reckase (1978) and Reckase (1978) compared fit of the one and three-parameter models to both empirical and simulated data. Their results indicated improvements in model fit when additional parameters (item discrimination and guessing) are considered in the model equations. Because the fit statistic used in these studies has unknown sampling properties, the conclusions must be viewed cautiously.

Hambleton and Cook (1978) made comparisons in fit of the one, two, and three-parameter models to simulated data. They found significant improvements in model fit for the more general models especially for short tests. The use of simulated data allows the researcher to compare "true" ability with ability estimated from the models. Although it is reasonable to expect improved fit to data when a model is less restrictive, unfortunately, studies involving simulated data do not provide a check on the adequacy of the model for describing

the real world.

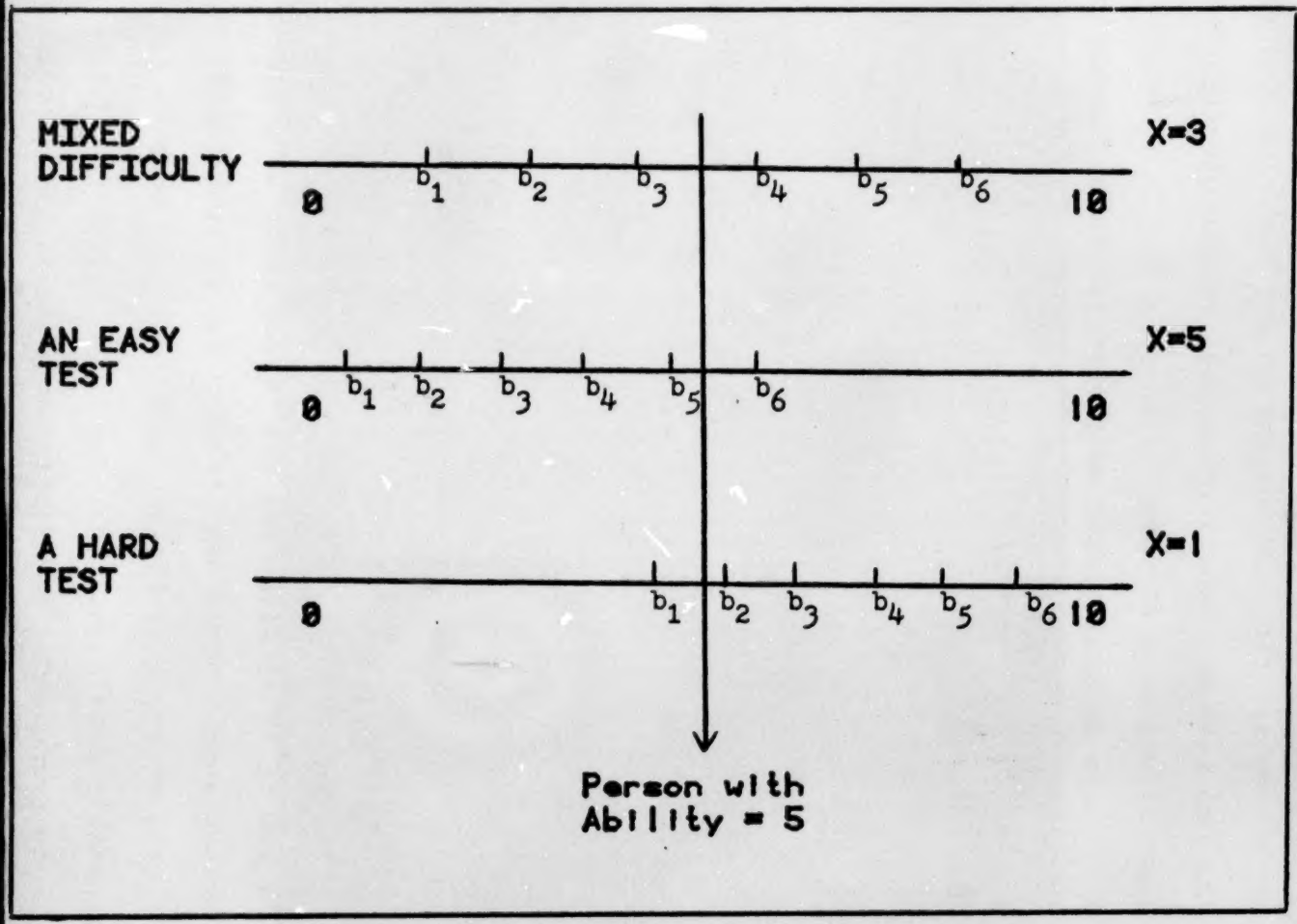
The study undertaken here differs from the previous research in a number of ways. First, the data utilized were empirical rather than simulated. Second, ability and item parameter estimates were determined using sophisticated computer routines developed for this purpose. Third, fit statistics with known sampling properties were utilized to make comparisons, and finally, twenty-five (25) datasets were employed, more than twice the number used in any previous comparative study. In addition, two alternative tests of fit were employed in the current research to help substantiate the results.

Concepts Utilized in Latent Trait Theory

This section presents some concepts underlying latent trait theory. A latent trait is the underlying trait or ability, inferred from an item or test score, which itself is not directly measurable. In classical testing methodology the number-right score is used to estimate the true score. The true score is actually a transformation of ability onto a number-right score scale. Ability estimates from latent trait theory are based on probabilistic models. When the difficulty level of an item is known it is possible to make inferences about ability from scores on single items in the following manner. Assume that an item has "G" difficulty. If an examinee obtains a correct response on the item, there is high probability that his ability is greater than or equal to G,

whereas if the examinee fails the item, it is highly probable that his ability is less than G. When ability estimates are made on a sufficient number of items, it is is possible to obtain an accurate measure of ability. Consistent estimates are found when test length is sufficiently long. Item difficulty is one of a number of possible item characteristics. In classical testing methodology number-right scores are obtained by summing correct responses across items without regard for the characteristics of the items. Depending on the specified latent trait model, one or more item characteristics are considered. Both models discussed here incorporate item difficulty. Figure 1 demonstrates the importance of item difficulty in obtaining ability estimates.

Figure 1
The Importance of Item Difficulty in Estimating Ability



Three hypothetical six item tests are shown in Figure 1: the top line of the figure depicts a test with items of mixed difficulty; the middle line shows an easy test; and the bottom line pictures a hard test. Item difficulty is measured on the same scale as ability which has arbitrarily been chosen to be 0 to 10. A single person, with ability five (5) on this scale, is illustrated in the figure. The number-right scores for this person on these tests, are respectively 3, 5, and 1. The conclusions drawn from each test utilizing classical testing methodology are very different. The latent trait ability estimate for this person would be the same regardless of which test was administered, since latent trait theory incorporates item difficulties into estimating ability. The closer item difficulty level is to examinee ability, the more important the item is for pinpointing ability. Because latent trait ability estimates do not depend on the specific sample of items in a test, they are considered to be objective. Similarly, latent trait ability estimates do not depend on the sample of persons tested.

The single item characteristic employed in obtaining ability estimates by the Rasch model is item difficulty. The three-parameter model utilizes item discrimination and "guessing" as well. Item discrimination is essentially used as a weight. The three-parameter model places emphasis on "good" (discriminating) items for estimating ability. Items can be selected which discriminate well at certain locations along the

ability continuum. The Rasch model assumes that all items are equally "good". In classical testing methodology this would mean that all item-total correlations are equal.

A "guessing" parameter is included in the three-parameter model to adjust estimates of ability at the low end of the scale where guessing is operable. If an examinee guesses, the probability of obtaining a correct response by chance alone is greater than zero. If the parameterswere not included, overestimates of ability would occur for low ability examinees. The Rasch model does not correct for guessing.

Two important assumptions for both models discussed here are unidimensionality and the equivalent assumption of local independence. Unidimensionality means that a test includes items which tap only a single underlying trait. Although there are multivariate extensions to latent trait models, these are not under consideration here. Local independence means that responses to items are statistically independent; the probability of a correct response on one item is not related to that on any other item. This means that for persons with identical ability, there is no correlation between item responses. If such a correlation were present, the test would be measuring more than one ability.

Model Descriptions

The latent trait models compared in this paper have the logistic form:

$$(1.0)\quad P = e^{x} / (1 + e^{x}) ,$$

where P is the probability of a correct response.
For the Rasch or one-parameter model, the probability function is:

$$(2.0)\quad P_g(\theta) = \frac{e^{D\bar{a}(\theta - b_g)}}{1 + e^{D\bar{a}(\theta - b_g)}} ,$$

and for the three-parameter model the probability of a correct response is given by:

$$(3.0)\quad P_g(\theta) = c_g + (1 - c_g) \frac{e^{Da_g(\theta - b_g)}}{1 + e^{Da_g(\theta - b_g)}} .$$

The function relating ability to the probability of a correct response is called an item characteristic curve (ICC). In equations (2.0) and (3.0) D is the constant 1.7 which is used to approximate the normal curve, since for all practical values of ability, the magnitude of the logistic function is within 1% of the normal ogive. Ability, θ, is measured on a standard score scale with practical values in the range -3 to +3. This scale may be linearly transformed to any arbitrary scale such as the LOGIT, a scale seen commonly in practice. Item difficulty, b, in the equations, is measured on the same scale as ability, with a practical range from -2 to +2. Item difficulty is the point on the ICC where the probability of a correct response is .5. Item discrimination and guessing parameters only apply to

the three-parameter model. Item discrimination, a, is measured on a scale ranging from 0 to +2. Negative values of discrimination are possible but of little interest here. Item discrimination is proportionate to the slope of the ICC at the point of its inflexion. In the Rasch ICC, equation (2.0), the average item discrimination is substituted for individual values since item discriminations are assumed equal. The guessing parameter in equation (3.0), c, ranges from 0 to 1.0 (the same range as P). This parameter forms the lower left asymptote to the item characteristic curve and represents the probability of success by chance alone. Practical limits for guessing are 0.1 to 0.5 depending on the number of choices, NCH. A good initial estimate of guessing is:

$$(4.0) \quad \hat{c}_g = (1 / NCH) - .05.$$

Since guessing is not assumed by the Rasch model, the parameter takes the zero value and does not appear in equation (2.0).

Two item characteristic curves for the one and three-parameter models are pictured respectively in Figures 2 and 3. The slopes of Rasch ICC's are identical and hence all one-parameter ICC's are parallel. The lower asymptotes of Rasch curves are all zero indicating no guessing. Three-parameter ICC's have different slopes and also vary in their lower asymptotes.

Figure 2
Three-Parameter Model ICCs

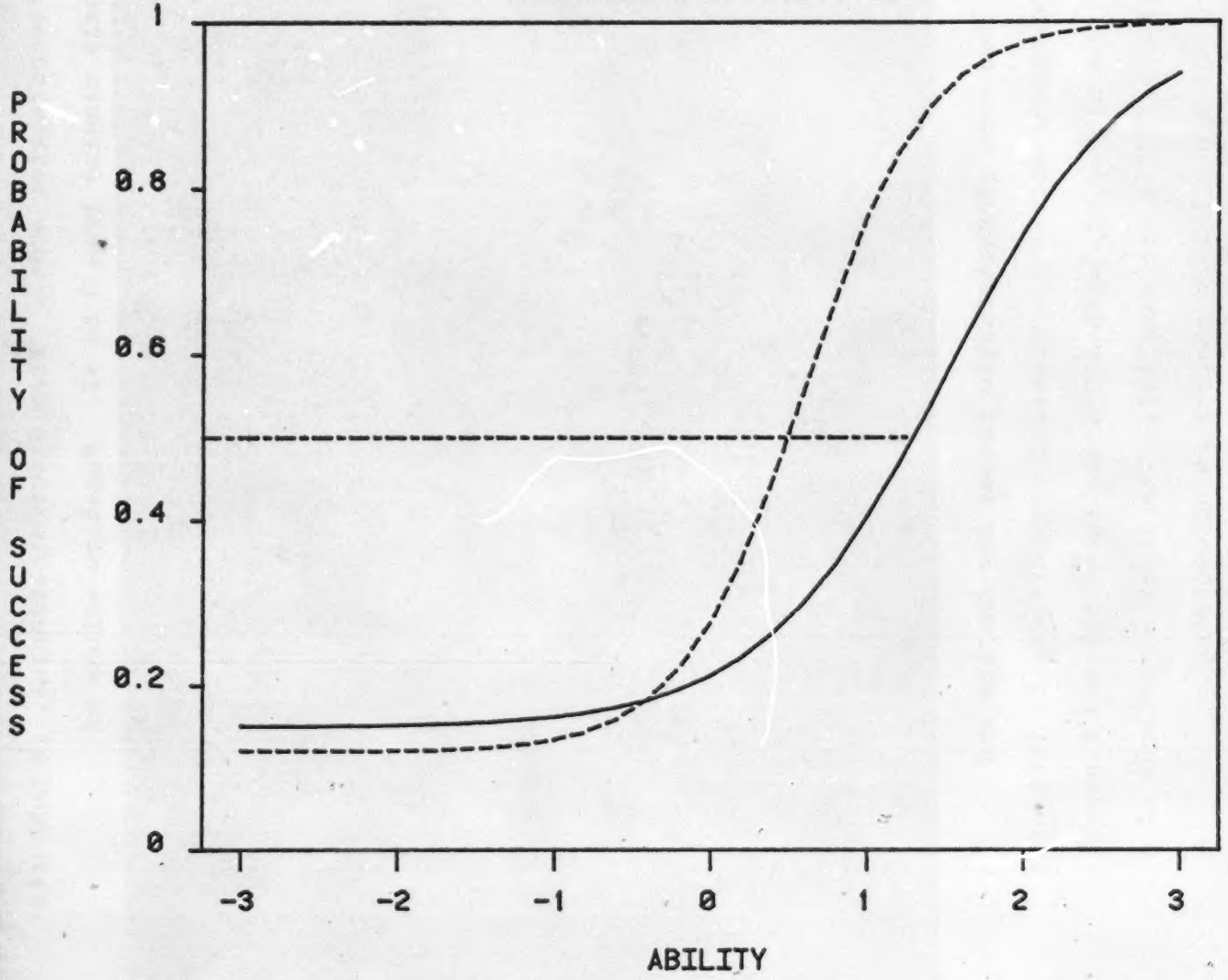
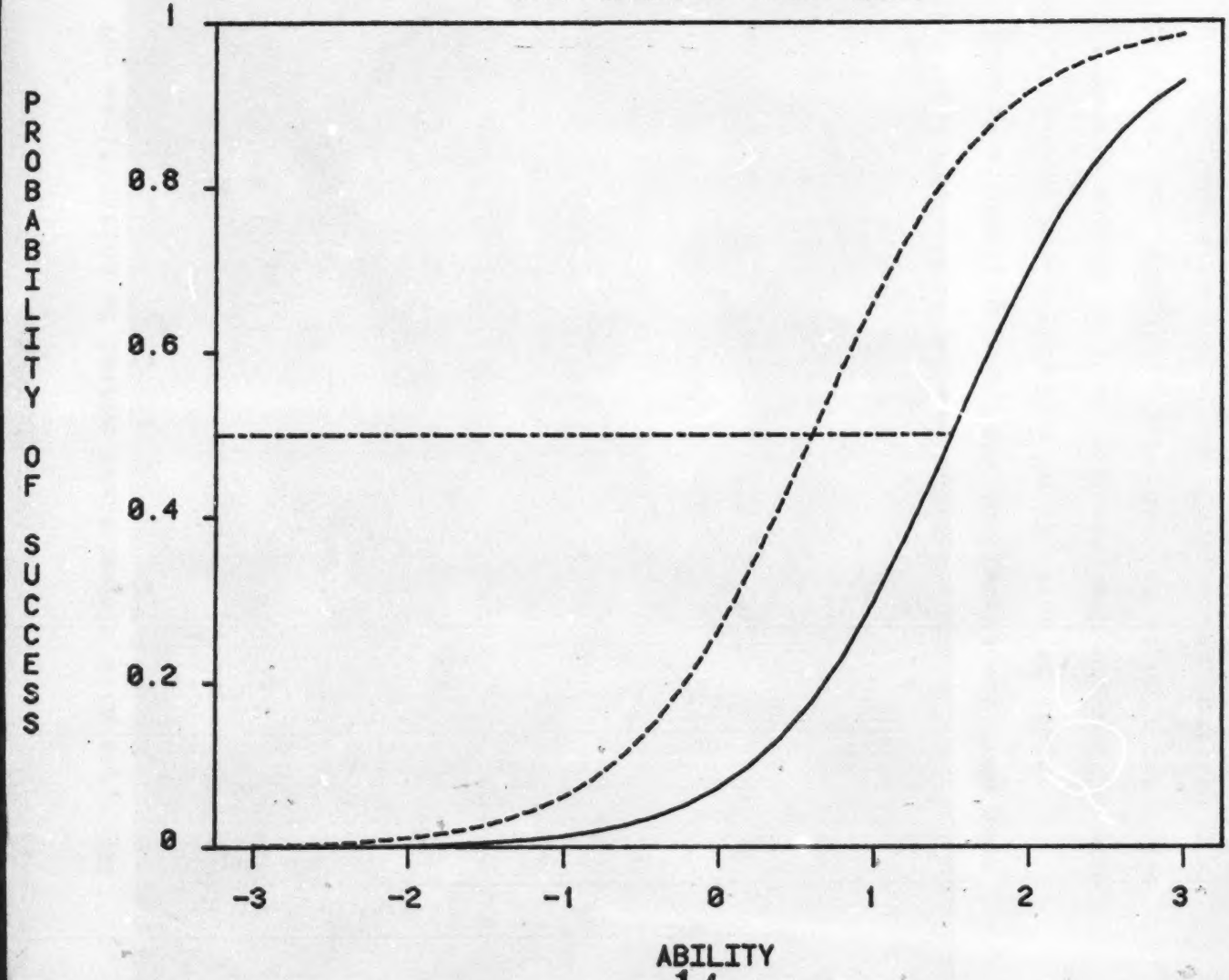


Figure 3
One-Parameter Model ICCs



METHODOLOGY

The verification of latent trait models with empirical data is complicated by the fact that both ability and item parameters are unknown. If ability could be directly measured, it would be elementary to check the adequacy of latent trait models. Instead we must make predictions based on a model and then attempt to verify the accuracy of these predictions using observable data. Estimated parameters are substituted for the unknown parameters in order to make predictions. In this study parameters are used to predict distributions of number-right scores based on each model and then the predicted distributions are compared to the observed number-correct score distributions. Since distributions are not assumed to be normal, non-parametric measures, including the chi square and Kolmogorov-Smirnov statistic, are used to make comparisons.

Twenty-five verbal and quantitative subtests from well recognized aptitude and achievement tests were selected for study. Data were obtained from test publishers and from data archives of persons doing research in the area of latent traits. A random sample of 1000 examiness was drawn from each test. When possible, examinees were sampled at only one grade or age level. Items were scored as right, wrong, omitted or "not reached" to conform to the requirements of the LOGIST computer program (Wood, Wingersky and Lord, 1976), which was used to obtain the latent trait item and ability estimates.

Data were then assessed on the degree that they met the assumptions of one or both models. Unidimensionality, an assumption of both models, was determined using a factor analytic criterion. Tetrachoric correlations, were used in the factor analysis because they are measure of association between two latent variables scored dichotomously. A principal components factoring solution without rotation was applied to the matrix of tetrachorics. The degree of unidimensionality was calculated as the ratio of eigenvalues between the first and second factors. This is equivalent to taking the variance ratio, a procedure which has been used extensively for this purpose. High values of this ratio indicate unidimensional tests. Low values suggest multidimensionality. The 25 tests were assigned ranks based on this ratio.

Two additional assumptions of the Rasch model were assessed: equality of item discriminations and no guessing. Although sample-free latent trait estimates of both of these quantities might have been used to make determinations about them, it was felt that these estimates (from the three-parameter model) would not be independent of the models and would produce biased results. Measures independent of the models were sought for testing these assumptions. Classical item difficulty and discrimination indices provided a suitable indicator of violations of these assumptions even though these are sample- based parameters.

The test for equality of item discriminationsused

classical point-biserials (item-total correlations). The mean point biserial for a test was calculated using a Fisher Z transformation to assure normality. Equal item-total correlations were defined as those falling within a .10 band around the mean. This confidence band was selected based on substantial experimentation with various ranges. The percentage of item-total correlations within this band was calculated as the indicator of violation of this model assumption. The 25 tests were then ranked on this criterion. A high rank indicated relatively equal item discriminations and a low rank indicated a wide range of item discriminations.

An appraisal of the extent of guessing proved to be impossible. The guessing parameter, as described earlier, is the probability of success on an item for examinees at the lowest end of the ability continuum. It forms the lower asymptote to the item characteristic curve. It is a parameter which describes the "guessability" of an item, not the tendency of an examinee to guess. In fact, inclusion of this parameter in the model suggests that low ability examinees guess correctly with a certain low, but non-zero probability. Item test regressions which are quite similar in shape to ICC's also display a non-zero asymptote. It would probably be inaccurate to assume that examinees do not guess when confronted with multiple choice items. Although these are substantial arguments for including the guessing parameter, a suitable measure for determining the amount of guessing, independent of the latent trait parameter estimate, could not be obtained.

Nevertheless a method for estimating the presence of guessing which employed classical item difficulty parameters was used. A description of the method follows with attention given both to understanding the rationale of the procedure and the shortcomings implicit in such an approach.

The following assumptions were made:

1. Guessing is a parameter of consequence only for hard items.
2. Guessing is operational only for low ability examinees.

Based on these assumptions hard items and low ability examinees were selected for study of this question. The following definitons of hard item and low ability examinee were used:

1. A hard item is one which two thirds of a sample scores incorrectly.
2. A low ability examinee is one in the lowest decile of a sample.

Unfortunately application of these definitions to each test produced strikingly different results. On some tests no hard items were found. On some tests the lowest decile had an average item difficulty of 40 percent and on other tests the average item difficulty for the low ability sample was only 10 percent. No suitable definition of hard item or low ability could be found to apply to all tests. Since each test represented different items and different people, there was no way to equate these definitions across tests. These problems highlight the inherent limitations of classical testing methodology which uses sample-dependent item parameters. The author received many alternative suggestions for estimating

guessing, but none were sample-free. Although results were generated using the above definitions, the author was quite aware that changing the definitions provided quite dissimilar results. Latent trait theory itself could provide the needed sample-free estimates for making this determination, but this would be assuming that the three-parameter model does fit data, an assumption that this research is attempting to prove. Additionally it should be noted that attempts to estimate guessing on samples of 1000 examinees proved disastrous: less than one-third of the guessing parameters were estimable in samples of this magnitude.

After ranking data for violations of model assumptions, item and ability parameters based on the assumptions of each model were estimated with the LOGIST computer program. Estimates of ability and item difficulty for each of the 25 datasets were found under the assumptions of the Rasch model. Ability, item difficulty, item discrimination and guessing estimates were found under the assumptions of the three-parameter model. By substituting these estimates for the true parameters, number-right score distributions were predicted from each model using the following procedure. When the ICC's for all items are identical, the distribution of number right scores can be expressed as the binomial distribution:

$$(5.0) \quad \phi (X \mid \theta) = \binom{N}{X} P^X Q^{N-X} \quad ,$$

where P is the probability of success on the items, N is the number of items, and X is the number of successes in N trials. In the more usual case, where the ICC for each item is different, the conditional distribution of X given ability can be expressed as a generalized binomial:

$$(6.0) \quad \phi (X \mid \theta) = \prod_{i}^{N} (Q_i + P_i) \quad ,$$

Although this distribution cannot be directly generated, it has a known mean and variance given by:

$$(7.0) \quad \mu_{X|\theta} = \sum_{i=1}^{N} P_i(\theta) \qquad \text{(mean)} ; \text{ and}$$

$$(8.0) \quad \sigma^2_{X|\theta} = \sum_{i=1}^{N} P_i Q_i \qquad \text{(variance)} .$$

The variance given by (8.0) is always less than the variance of the binomial. Utilizing (7.0) and (8.0) it is possible to determine the distribution of X, number right score, for any given level of ability. Rather than estimating the distribution for each individual ability level in the study, abilities were grouped into 17 divisions each .5 wide in the range -3.25 to 3.25. The midpoint of each grouping was used to predict the number-right score distribution for the group. These frequencies were summed across the 17 levels of ability and converted to frequency counts of examinees at each number-right score level. These were then compared to frequency counts in the observed score distributions. The number-right score is defined as:

$$(9.0) \quad X = \sum_{g=1}^{n} u_g \quad ,$$

where u = 1 for a correct response and zero otherwise.

Several techniques were used to compare the predicted with the observed score distributions. Since no assumptions of normality were made, non-parametric techniques were applied. The chi square statistic provided a means for comparing distributions of grouped raw scores, where three or four score levels were combined into a group. Additional regrouping was accomplished when cell sizes were under 5. When N is sufficiently large, most data is rejected by the chi square criterion. This was not a consideration in this study since the statistics were used to compare models relatively rather than as an absolute criterion for rejection. The degrees of freedom for the chi square test is based on the number of groups. The Kolmogorov-Smirnov (K-S) statistic was used to compare ungrouped distributions (single raw score per group). The K-S measure can detect differences at any point along a continuum and does not depend on the concept of degrees of freedom. A third measure of comparison between observed and predicted distributions was a graphic technique. This method highlighted the location of discrepancies along the curves and could be used to examine data in terms of the shapes of the distributions (normal, uniform, skewed). Fit statistics were then used to explore model fit to data in which violations of model assumptions occurred. Chi square and K-S fit statistics were correlated with indicators and ranked measures of violation of model assumptions using Pearson and Spearman

correlation techniques.

During the parameter estimation stage of the design, data were collected on the computer resources needed to estimate parameters under the assumptions of each model. Both computer times and costs were recorded to provide a guide for persons selecting models.

RESULTS

Table I presents summary statistics for the twenty-five datasets. The table includes the dataset mnemonic, test length, indicators of violations of three model assumptions, and Kolmogorov-Smirnov (K-S) fit statistics for each model. The chi square statistics were highly correlated with K-S statistics and are consequently not reported. Overall, the three-parameter model fit data slightly better on the average than the one-parameter model. The average K-S statistic for the Rasch model was 1.3, and for the three-parameter model 1.2. As an absolute criterion, five tests were rejected with the K-S statistic at the .05 level of probability and one test at the .10 level. There was no relationship between test length and magnitude of the K-S statistic, although this relationship was seen for chi square statistics (longer tests were rejected more often).

Table II presents measures of association between indicators of violations of model assumptions and fit. Both Pearson and Spearman rank order correlations are included. The results suggest that unidimensionality was an important indicator of model fit. The correlations between degree of unidimensionality and model fit are significant at the .05 level. The correlations for model fit with indicators of equality of item discrimination and guessing were not significant. Fit to both models was slightly better when a broad range of item discriminations characterized the data rather than equal item discriminations. The guessing results

TABLE I

Summary Statistics for Twenty-five Datasets

Dataset	Length	Degree of unidimensionality[c]	% Equal item discriminations[d]	Extent of Guessing[e]	K-S 3[f]	K-S 1[f]
Stan.W.S.	50	9.08	72%	a	1.07	1.15
Stan.Read.	71	3.41	61	27%	3.00	3.38
SAT Verb.	85	4.65	78	36	b	b
CTBS Math	48	11.59	73	0	b	b
CTBS Vocab.	40	14.96	70	25	b	b
Stan. Vocab.	50	7.23	82	56	b	b
Stan. Sci.	60	6.43	82	67	b	b
Stan. Lang.	80	4.94	71	42	1.52	1.62
Stan.Soc.Stud.	54	5.81	67	75	1.26	1.23*
Stan. Listen.	50	8.80	84	50	1.12	.90*
Stan. Spell.	60	7.58	72	33	1.18	1.06*
Stan. Math App	40	6.34	88	50	.87	.94
Stan. Math Con	35	5.46	74	67	1.33	1.45
Crit.Ref.Test	16	5.44	63	a	1.45	1.62
SAT Verb.	85	4.32	78	58	1.48	1.49
Reading 1	70	3.33	70	86	1.15	1.14*
Reading 2	70	4.97	69	50	1.10	1.16
Reading 3	70	5.12	59	67	1.34	1.14*
CAT Verb.	40	5.19	50	55	1.33	1.25*
ITBS Verb.	38	7.65	74	60	1.31	1.15*
CAT Comp.	45	5.84	67	50	1.06	1.09
ITBS Comp.	68	5.34	62	67	.99	.98*
Rasch Test 1	34	6.65	50	100	1.01	1.01
Rasch Test 2	30	3.86	87	a	1.22	1.21*
Rasch Test 3	39	5.99	77	100	1.00	1.11

[a] No difficult items on test

[b] Chi square statistics found for these tests

[*] Better fit for one-parameter model

[c] Ratio of first to second latent roots

[d] Percent of item-total correlations within a band $\bar{r} \pm .1$

[e] Percent of guessing on hard items (items answered correctly by only 1/3rd of examinees) for lowest decile of examinees

[f] Kolmogorov-Smirnov statistic

TABLE II

Pearson and Rank order Correlations Between Kolmogorov-Smirnov Fit Statistics for the One and Three-parameter models and Three indicators of Congruence with Model Assumptions

Indicator	Kolmogorov-Smirnov Fit Statistic	
	3-parameter model	1-parameter model
Unidimensionality[a]		
Pearson	-.468*	-.555*
Spearman	-.474*	-.544*
Equality of Item Discriminations[a]		
Pearson	-.196	-.158
Spearman	-.238	-.221
Extent of Guessing[b]		
Pearson	-.319	-.190
Spearman	-.118	0.000

[a] n=20 for Spearman, n=16 for Pearson

[b] Only includes those tests for which there were minimally 4 hard items

* significant at .05 level

TABLE III

Computer Times and Costs for Estimation of Latent[a] Trait Parameters under the One and Three-parameter Model Assumptions

	3-parameter model	1-parameter model
Average time	92.4 sec.	33.7 sec.
Average cost	$35.12	$12.80

[a] computed on a Control Data Corporation CYBER-175 computer under the NOS 1.3 operating system

used the rather arbitrary criterion discussed above. Altering this criterion produced correlations varying in both sign and magnitude and therefore no conclusions are reached concerning fit of data to the Rasch model when there is guessing.

Figures 4 and 5 are graphic comparisions of the predicted and observed distributions. The first figure shows a uniform distribution; the second figure shows a skewed distribution. The figures typify results obtained on 25 datasets. The following observations apply to both models. The largest chi square and K-S statistics occurred when the observed distributions were very irregular in shape. Uniform distributions were frequently rejected by the chi square criterion when tests were long, but for distributions of skewed form, such as those in the second figure, short tests were more frequently rejected. In addition, when distributions were skewed (either left or right), estimated score distributions tended to be shifted to the right, indicating an underestimation of low ability examinees and an overestimation of high ability examinees by both models.

Table III presents computer times and costs for parameter estimation under the assumptions each model. Three-parameter estimates of item and ability parameters require three times the resources needed by the Rasch model, although it should be noted that when item parameters are known, the cost for estimating ability under the two models is identical.

Figure 4
Observed and Expected Score Frequencies

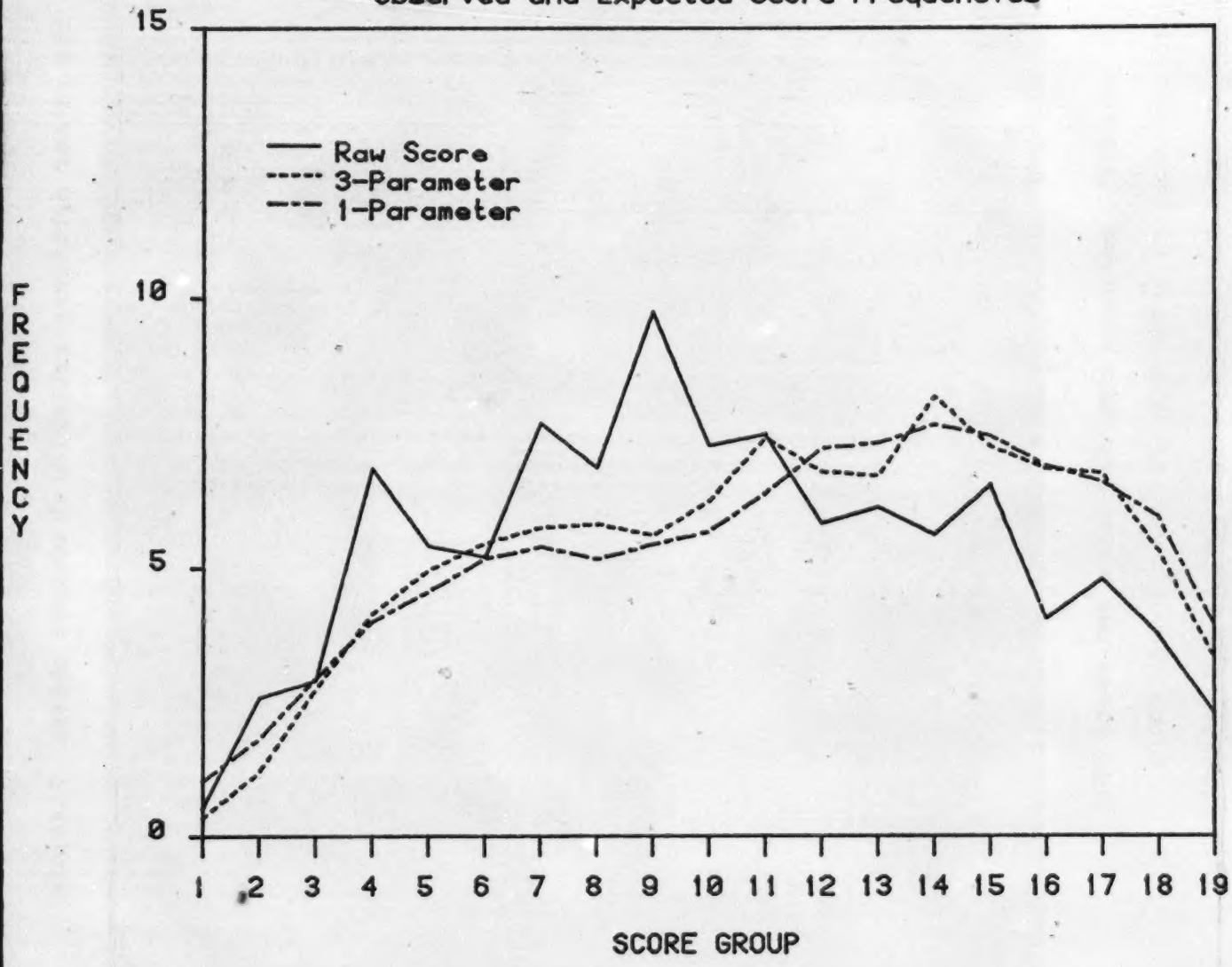
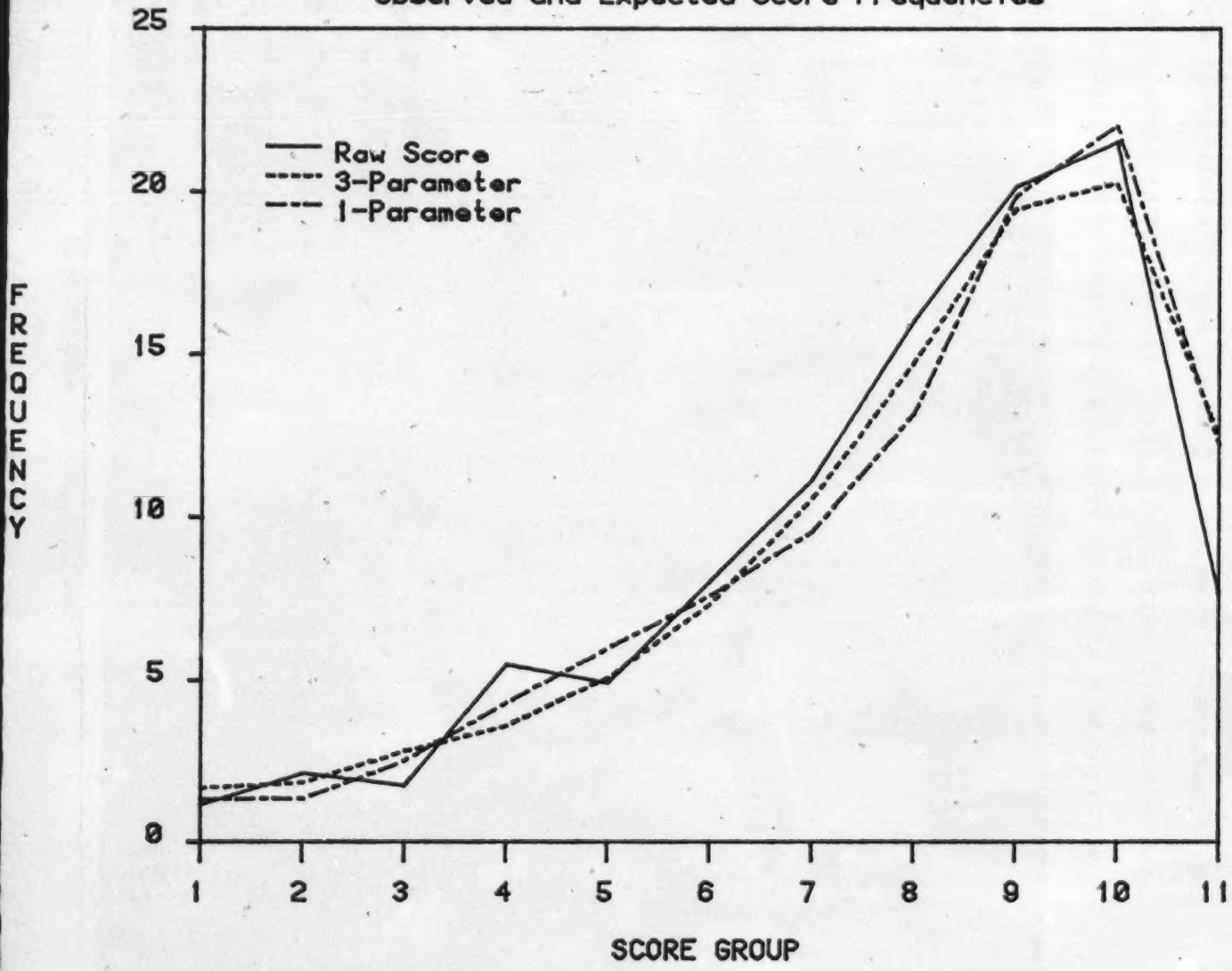


Figure 5
Observed and Expected Score Frequencies



DISCUSSION OF THE RESULTS

The results of this study suggest that for purposes of estimating ability by latent trait methods, the Rasch model compares favorably with the three-parameter logistic model. Using estimated parameters to make predictions about number-correct score distributions, those predicted by the Rasch model fit the observed distribution of number-correct scores as well as the distributions predicted by the three-parameter model. This suggests that the Rasch model is quite robust to violations of its underlying assumptions and can therefore be applied in a wide variety of situations.

The single model assumption related to fit was that of unidimensionality. Not only was fit for multidimensional tests quite poor, but it was also difficult to obtain parameter estimates on such tests. This suggests that it is advisable to form unidimensional item subsets from multi-trait tests before obtaining ability estimates.

The results of the study concerning equality of item discriminations are contrary to those found in a simulation study by Dinero and Haertel (1977). The item discriminations in the 25 datasets of the current study were consistently high and displayed small variances around the mean. It is suspected that the ranges of discriminations in the current study may be much smaller than those in the simulated data produced by Dinero and Haertel. One dataset was selected to probe the

discrepancy in results between the two studies. Two short 20 item tests were constructed from an 80 item test; one composed of relatively equal item discriminations and the other composed of items with the highest and lowest item discriminations from the original test. The two tests were then fit to the one and three-parameter models. In this case, the test with the extremely broad range of item discriminations did not fit the Rasch model. From these results it can be concluded that much empirical data may have a sufficiently narrow ranges of item discrimination to be acceptable for Rasch model procedures. Further simulation studies could demonstrate the variance of item discriminations that can be tolerated by the Rasch model.

It was impossible to reach any conclusions on the effect of violation of the no guessing assumption and fit to the Rasch model. Simulated data would also prove useful for exploration of this question. Using simulation techniques, data could be generated under the assumptions of the three-parameter model, but tested for fit to the Rasch model. The amount of guessing would be varied to determine how robust the Rasch model is to violations of this assumption. Although there are many indicators that ICC's are characterized by non-zero lower asymptotes, only one-third of guessing parameters were estimable by the LOGIST program for samples of 1000 examinees. Since three-parameter estimation is costly by comparison to Rasch estimation, it may be more practical to set the lower asymptote to some reasonable value without attempting more refined estimation.

This study used estimated parameters to predict number-right score distributions, although in previous research weighted number-right score distributions were used. There is no evidence in the latent trait literature for favoring either one of these criteria. This author felt that use of the weighted number-right score would produce results biased in favor of the three-parameter model, hence the more conservative approach was taken.

Other research indicates that the three-paramater model may be preferable for test development, test equating, and studies of item bias, but for estimating ability, this study shows that the Rasch model produces estimates of ability that rival those produced by the three-parameter model. Further, the study demonstrates that Rasch parameter estimates are obtained at costs significantly reduced from those of the three-parameter estimates.

REFERENCES

Dinero, T.E. and Haertel, E. Applicability of the Rasch model with varying item discriminations. Applied Psychological Measurement, 1977, 1, 581-592.

Hambleton, R.K. An empirical investigation of the Rasch test theory model. Unpublished doctoral dissertation, University of Toronto, 1969.

Hambleton, R.K. and Cook, L.L. Some results on the robustness of latent trait models. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Hambleton, R.K. and Traub, R.E. Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 1973, 26, 195-211.

Koch, B.R. and Reckase. M.D. A live tailored testing comparison of the one- and three-parameter logistic models. Paper presented at the National Council on Measurement in Education annual meeting, Toronto, 1978.

Reckase, M.D. A comparison of the one- and three-parameter logistic models for item calibration. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Wood, R.L., Wingersky, M.S., and Lord, F.M. LOGIST: A computer program for estimating ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, NJ: Educational Testing Service, 1976.

TECHNICAL NOTES

This paper was produced on a DEC-20 computer with the EMACS Editor written in TECO by Richard Stallman at the Artificial Intelligence Laboratory, MIT. The final copy was printed on an Anderson-Jacobsen 841 "golfball" style printing terminal. All figures were made with the Tektronix Plot 10 Easy Graphing package on a Tektronix 4013 terminal with attached flatbed plotter. My appreciation is extended to Wellesley College for providing these excellent facilities.

Data analysis was accomplished on the Control Data CYBER-70 computer at the University of Massachusetts, Amherst. In addition to the LOGIST program, many SPSS utilities were accessed. All other computer programs were developed by the author including a computerized Latent Trait bibliography system and a time-sharing interface for LOGIST. My thanks is offerred to the University of Massachusetts computing staff, especially to the operators who assisted in mounting 50 magnetic tapes used in the study.